PRIS at TREC2012 KBA Track

Yan Li, Zhaozhao Wang, Baojin Yu, Yong Zhang, Ruiyang Luo,
Weiran Xu, Guang Chen, Jun Guo
School of Information and Communication Engineering,
Beijing University of Posts and Telecommunications
Beijing, P.R. China, 100876
buptliyan@gmail.com

Abstract

Our system to KBA Track at TREC2012 is described in this paper, which includes preprocessing, index building, relevance feedback and similarity calculation. In particular, the Jaccard coefficient was applied to calculate the similarities between documents. We also show the evaluation results for our team and the comparison with the best and median evaluations.

1. Introduction

Knowledge Base Acceleration (KBA) seeks to help humans expand knowledge bases like Wikipedia by automatically recommending edits based on incoming content streams. For our first year in TREC, we are evaluating systems on a single, simple task called cumulative citation recommendation: filter a stream of content for information that should be linked from a given Wikipedia page or an specific entity.

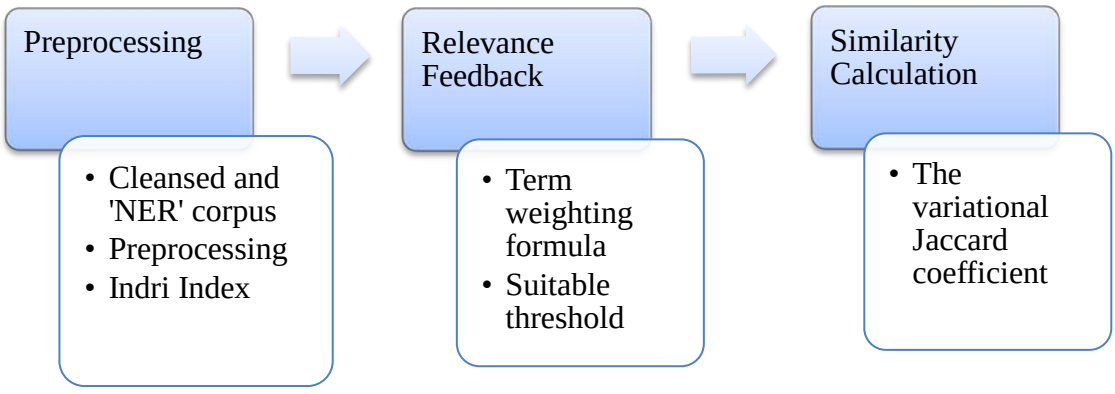

Figure 1. The framework of KBA system

Figure 1 shows the framework of our KBA system. First of all, we focused on the “cleansed” and “NER” part of the corpus. Preprocessing filtered out the useless documents and information and built the Indri index of the remain corpus. Secondly, the relevance feedback was conducted to

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
NOV 2012		**00-00-2012 to 00-00-2012**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
PRIS at TREC2012 KBA Track	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
Beijing University of Posts and Telecommunications,School of Information and Communication Engineering,Beijing, P.R. China, 100876,	

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

Presented at the Twenty-First Text REtrieval Conference (TREC 2012) held in Gaithersburg, Maryland, November 6-9, 2012. The conference was co-sponsored by the National Institute of Standards and Technology (NIST) the Defense Advanced Research Projects Agency (DARPA) and the Advanced Research and Development Activity (ARDA). U.S. Government or Federal Rights License

14. ABSTRACT

Our system to KBA Track at TREC2012 is described in this paper, which includes preprocessing, index building, relevance feedback and similarity calculation. In particular, the Jaccard coefficient was applied to calculate the similarities between documents. We also show the evaluation results for our team and the comparison with the best and median evaluations.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**5**	

expand the query information. Three expanded terms were generated for each entity. We used these terms to query the index and obtained an initial candidate for the relevant documents to be recommended. Finally we utilized an variation of the Jaccard coefficient to calculate the similarities between documents and generate the final recommended documents according to a threshold.

2. Preprocessing and Index Building

To fulfill the succeeding algorithm, we need to preprocess the original corpus and build a index for the retrieval system.

After deciphering the corpus using a standard gpg and XZ decompression, we get the original data collected from Wikipedia. The corpus has been split into three components: linking, social and news. We only focused on documents labeled with 'cleansed' & 'ner', and extracted essential part for index building. Then some text processing procedures were executed for these documents:

- Non-English text deletion
- Lowercasing the capital letters
- Removing the external linking inside the text
- Abbreviation expansion
- Removing useless punctuations

We converted the documents into the “trectext” format used by Indri toolset for building index. Besides the text itself, we kept the information of “DOCNO”, “stream_id” and “Time”. We used a simple stop word list to help Indri exclude useless words. In addition, the Porter algorithm was used for the stemming task.

3. Relevance Feedback

KBA uses entities as filter topics for this year’s CCR task. However, it is not enough to retrieve the index just according to a single entity name. In order to get more information about the topic, we expanded the topic entity utilizing two kinds of profiles. One is the Wikipedia page of the entity and another is the annotation set provided by TREC. From the annotation, we picked out documents labeled with either ‘R’ (Relevant) or ‘C’ (Central) for each entity.

After that we used the following formula to calculate the weight of each term:

$$P_{ml}(t|M_d) = \frac{tf_{(t,d)}}{dl_d} \quad (1)$$

$$P_{avg}(t) = \frac{\sum_{d_{(t\in d)}} P_{ml}(t|M_d)}{df_t} \quad (2)$$

where $tf_{(t,d)}$ is the raw term frequency of term t in document d, dl_d is the total number of tokens in document d, dft is the document frequency of t and $P_{avg}(t)$ is the weight of each word. Then we set a threshold to choose the top three words as the final expanded queries. Besides the initial entity topic, these terms were queried searching the index to find out the candidate similar documents.

4. Similarity Calculation

For the purpose of generating final recommended documents from the candidate above, we utilized the Jaccard coefficient to calculate the similarity between candidate documents and the original Wikipedia page for each topic entity. The Jaccard similarity coefficient is a statistic used for comparing the similarity and diversity of sample sets. The Jaccard coefficient measures similarity between sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets. We used an variation of the traditional Jaccard formula for our specific task showing as follows:

$$R_{wiki,d} = \frac{\sum_{t \in |wiki \cap d|} tf_{wiki}(t) * tf_d(t)}{\sum_{t \in |wiki \cap d|} [tf_{wiki}^2(t) + tf_d^2(t)]} \quad (3)$$

where *wiki* and *d* stands for Wikipedia page and candidate document respectively. $tf(t)$ means the term frequency of *t*. The calculated Jaccard coefficient should be multiplied by 1000 as the final confidence score for each candidate. We then compared the confidence score with the similarity threshold: if the coefficient is larger than the threshold, the document is recommended. The threshold is actually set from 400 to 1000.

5. Evaluation Results

We have submitted up to 7 runs for this year's task. Due to the limited space, we only show the best results of us and the comparison of others.

Table 1 and Figure 2 shows the Precision, Recall, F1 and Scaled Utility of our run. It can be seen that F1 measure increases when the cutoff goes down and arrives peak at 400 cutoff, whereas the Scaled Utility shows an inverse trend.

Table 1. Average performance of the PRIS run

cutoff	Precision	Recall	F1	Scaled Utility
0	0.267298	0.05809	0.067795	0.250206
100	0.267298	0.05809	0.067795	0.250206
200	0.267298	0.05809	0.067795	0.250206
300	0.267298	0.05809	0.067795	0.250206
400	0.267298	0.05809	0.067795	0.250206
500	0.210405	0.02739	0.041212	0.292443
600	0.056385	0.005454	0.008731	0.304233
700	0.03046	0.003025	0.005293	0.319105
800	0	0	0	0.325258
900	0	0	0	0.33285

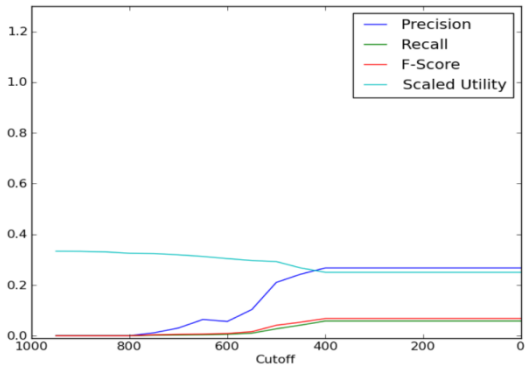

Figure 2. Average performance of the PRIS run

Table 2. Comparison with the Best, Median and Mean F1 measure on cutoff 400

URL name	PRIS	Best	Median	Mean
Aharon_Barak	0.1163	0.3841	0.1909	0.1664
Alex_Kapranos	0	0.4298	0.2706	0.2263
Alexander_McCall_Smith	0.0832	0.3963	0.1955	0.1593
Annie_Laurie_Gaylor	0.0233	0.5021	0.3046	0.2304
Basic_Element (company)	0.0952	0.8497	0.1670	0.2714
Basic_Element (music_group)	0.0104	0.8483	0.0757	0.1238
Bill_Coen	0.0769	0.4375	0.1984	0.1709
Boris_Berezovsky_(businessman)	0.0015	0.5371	0.4859	0.3503
Boris_Berezovsky_(pianist)	0	0.5714	0.0045	0.0369
Charlie_Savage	0.0202	0.6846	0.1135	0.1339
Darren_Rowse	0.1505	0.3271	0.1910	0.1676
Douglas_Carswell	0	0.5562	0.1352	0.1286
Frederick_M._Lawrence	0.1818	0.7027	0.2684	0.2621
Ikuhisa_Minowa	0	0.5860	0.5229	0.3749
James_McCartney	0.0293	0.5757	0.2637	0.2275
Jim_Steyer	0.0556	0.7419	0.4599	0.3296
Lisa_Bloom	0.0566	0.6341	0.1302	0.1524
Lovebug_Starski	0	0.2462	0.1176	0.0913
Mario_Garnero	0.4930	0.9211	0.7741	0.6095
Masaru_Emoto	0.1091	0.2	0.1014	0.0843
Nassim_Nicholas_Taleb	0.0056	0.4747	0.3143	0.2578
Rodrigo_Pimentel	0.0390	0.5385	0.0751	0.1168
Roustam_Tariko	0.0408	0.4982	0.3634	0.2786
Ruth_Rendell	0.0132	0.4430	0.3357	0.2304
Satoshi_Ishii	0.0061	0.6556	0.4239	0.3266
Vladimir_Potanin	0.0552	0.7508	0.2556	0.2580
William_Cohen	0	0.3484	0.0816	0.0815
William_D._Cohan	0.0529	0.6538	0.3458	0.3041
William_H._Gates,_Sr	0.3039	0.3943	0.1803	0.1491
average	0.0678	0.4263	0.2506	0.2066

Table 2 shows the comparison between our run and the best, median and mean results on F1 measure at cutoff 400. We can conclude that the F1 measures of two entities (Masaru_Emoto and William_H.Gates,_Sr) are higher than the median and mean results; the F1 measures of four entities (Aharon_Barak, Darren_Rowse, Frederick_M._Lawrence and Mario_Garnero) are comparable to the median and mean while the results of remaining entities are lower than average.

The average F value is 0.0678 while the average median and mean is 0.2506 and 0.2066 respectively, which means that there is still a large room for improvement.

6. References

[1] http:// http://trec-kba.org/kba-ccr-2012.shtml /

[2] Si Li et.al, PRIS at 2009 Relevance Feedback track: Experiments in Language Model for Relevance Feedback. In proceedings of the 18th TREC 2009.